Filters that Fight Back

paulgraham.com · Paul Graham · 2003-08 · source

|
August 2003

We may be able to improve the accuracy of Bayesian spam filters by having them follow links to see what's waiting at the other end. Richard Jowsey of death2spam now does this in borderline cases, and reports that it works well.

Why only do it in borderline cases? And why only do it once?

As I mentioned in Will Filters Kill Spam?, following all the urls in a spam would have an amusing side-effect. If popular email clients did this in order to filter spam, the spammer's servers would take a serious pounding. The more I think about this, the better an idea it seems. This isn't just amusing; it would be hard to imagine a more perfectly targeted counterattack on spammers.

So I'd like to suggest an additional feature to those working on spam filters: a "punish" mode which, if turned on, would spider every url in a suspected spam n times, where n could be set by the user. [1]

As many people have noted, one of the problems with the current email system is that it's too passive. It does whatever you tell it. So far all the suggestions for fixing the problem seem to involve new protocols. This one wouldn't.

If widely used, auto-retrieving spam filters would make the email system rebound. The huge volume of the spam, which has so far worked in the spammer's favor,

would now work against him, like a branch snapping back in his face. Auto-retrieving spam filters would drive the spammer's costs up, and his sales down: his bandwidth usage would go through the roof, and his servers would grind to a halt under the load, which would make them unavailable to the people who would have responded to the spam.

Pump out a million emails an hour, get a million hits an hour on your servers.

We would want to ensure that this is only done to suspected spams. As a rule, any url sent to millions of people is likely to be a spam url, so submitting every http request in every email would work fine nearly all the time. But there are a few cases where this isn't true: the urls at the bottom of mails sent from free email services like Yahoo Mail and Hotmail, for example.

To protect such sites, and to prevent abuse, auto-retrieval should be combined with blacklists of spamvertised sites. Only sites on a blacklist would get crawled, and sites would be blacklisted only after being inspected by humans. The lifetime of a spam must be several hours at least, so it should be easy to update such a list in time to interfere with a spam promoting a new site. [2]

High-volume auto-retrieval would only be practical for users on high-bandwidth connections, but there are enough of those to cause spammers serious trouble. Indeed, this solution neatly mirrors the problem. The problem with spam is that in order to reach a few gullible people the spammer sends mail to everyone. The non-gullible recipients are merely collateral damage. But the non-gullible majority won't stop getting spam until they can stop (or threaten to stop) the gullible from responding to it. Auto-retrieving spam filters offer

them a way to do this.

Would that kill spam? Not quite. The biggest spammers could probably protect their servers against auto-retrieving filters. However, the easiest and cheapest way for them to do it would be to include working unsubscribe links in their mails. And this would be a necessity for smaller fry, and for "legitimate" sites that hired spammers to promote them. So if auto-retrieving filters became widespread, they'd become auto-unsubscribing filters.

In this scenario, spam would, like OS crashes, viruses, and popups, become one of those plagues that only afflict people who don't bother to use the right software.

Notes

[1] Auto-retrieving filters will have to follow redirects, and should in some cases (e.g. a page that just says "click here") follow more than one level of links. Make sure too that the http requests are indistinguishable from those of popular Web browsers, including the order and referrer.

If the response doesn't come back within x amount of time, default to some fairly high spam probability.

Instead of making n constant, it might be a good idea to make it a function of the number of spams that have been seen mentioning the site. This would add a further level of protection against abuse and accidents.

[2] The original version of this article used the term "whitelist" instead of "blacklist". Though they were to work like blacklists, I preferred to call them whitelists because it might make them less vulnerable to legal attack. This just seems to have confused readers, though.

There should probably be multiple blacklists. A single point of failure would be vulnerable both to attack and abuse.

Thanks to Brian Burton, Bill Yerazunis, Dan Giffin, Eric Raymond, and Richard Jowsey for reading drafts of this.

|